The current genomic landscape of western South America: Andes, Amazonia and Pacific Coast

Barbieri, Chiara ; Barquera, Rodrigo ; Arias, Leonardo ; Sandoval, José R ; Acosta, Oscar ; Zurita, Camilo ; Aguilar-Campos, Abraham ; Tito-Álvarez, Ana M ; Serrano-Osuna, Ricardo ; Gray, Russell D ; Heggarty, Paul ; Shimizu, Kentaro K ; Fujita, Ricardo ; Stoneking, Mark ; Pugach, Irina ; Fehren-Schmitz, Lars

Abstract: Studies of Native South American genetic diversity have helped to shed light on the peopling and differentiation of the continent, but available data are sparse for the major ecogeographic domains. These include the Pacific Coast, a potential early migration route; the Andes, home to the most expansive complex societies and to one of the most widely-spoken indigenous language families of the continent (Quechua); and Amazonia, with its understudied population structure and rich cultural diversity. Here we explore the genetic structure of 176 individuals from these three domains, genotyped with the Affymetrix Human Origins array. We infer multiple sources of ancestry within the Native American ancestry component; one with clear predominance on the Coast and in the Andes, and at least two distinct substrates in neighboring Amazonia, including a previously undetected ancestry characteristic of northern Ecuador and Colombia. Amazonian populations are also involved in recent gene-flow with each other and across ecogeographic domains, which does not accord with the traditional view of small, isolated groups. Long-distance genetic connections between speakers of the same language family suggest that indigenous languages here were spread not by cultural contact alone. Finally, Native American populations admixed with post-Columbian European and African sources at different times, with few cases of prolonged isolation. With our results we emphasize the importance of including under-studied regions of the continent in high-resolution genetic studies, and we illustrate the potential of SNP chip arrays for informative regional-scale analysis.

DOI: https://doi.org/10.1093/molbev/msz174

Paul; Shimizu, Kentaro K; Fujita, Ricardo; Stoneking, Mark; Pugach, Irina; Fehren-Schmitz, Lars (2019). The current genomic landscape of western South America: Andes, Amazonia and Pacific Coast. Molecular Biology and Evolution, 36(12):2698-2713.
DOI: https://doi.org/10.1093/molbev/msz174

Article: Discoveries

The current genomic landscape of western South America: Andes, Amazonia and Pacific Coast

Chiara Barbieri[a,b]*, Rodrigo Barquera[c], Leonardo Arias[d], José R. Sandoval[e], Oscar Acosta[e], Camilo Zurita[f,g], Abraham Aguilar-Campos[h], Ana M. Tito-Álvarez[i], Ricardo Serrano-Osuna[h], Russell Gray[a], Fabrizio Mafessoni[d], Paul Heggarty[a], Kentaro K. Shimizu[b], Ricardo Fujita[e], Mark Stoneking[d], Irina Pugach[d], Lars Fehren-Schmitz[j,k]

[a] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, 07745, Germany

[b] Department of Evolutionary Biology and Environmental Studies, University of Zurich, 8057, Switzerland

[c] Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, 07745, Germany

[d] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany

[e] Centro de Investigación de Genética y Biología Molecular (CIGBM) Universidad de San Martín de Porres, Lima, 15024, Peru

[f] Cátedra de Inmunología, Facultad de Medicina, Universidad Central del Ecuador, Quito, 170136, Ecuador

[g] Unidad de Investigaciones en Biomedicina, Zurita&Zurita Laboratorios, Quito, 170104, Ecuador

[h] Clinical Laboratory, Unidad Médica de Alta Especialidad (UMAE) # 2, Instituto Mexicano del Seguro Social (IMSS), Ciudad Obregón, 85130, Sonora, Mexico.
[i] Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Quito, 170125, Ecuador.

[j] UCSC Paleogenomics, Department of Anthropology, University of California, Santa Cruz, CA 95064, USA

[k] Genomics Institute, University of California, Santa Cruz, CA 95064, USA

* To whom correspondence may be addressed. Chiara Barbieri, University of Zurich. Email: barbieri.chiara@gmail.com

Abstract

Studies of Native South American genetic diversity have helped to shed light on the peopling and differentiation of the continent, but available data are sparse for the major ecogeographic domains. These include the Pacific Coast, a potential early migration route; the Andes, home to the most expansive complex societies and to one of the most widely-spoken indigenous language families of the continent (Quechua); and Amazonia, with its understudied population structure and rich cultural diversity. Here we explore the genetic structure of 176 individuals from these three domains, genotyped with the Affymetrix Human Origins array. We infer multiple sources of ancestry within the Native American ancestry component; one with clear predominance on the Coast and in the Andes, and at least two distinct substrates in neighboring Amazonia, including a previously undetected ancestry characteristic of northern Ecuador and Colombia. Amazonian populations are also involved in recent gene-flow with each other and across ecogeographic domains, which does not accord with the traditional view of small, isolated groups. Long-distance genetic connections between speakers of the same language family suggest that indigenous languages here were spread not by cultural contact alone. Finally, Native American populations admixed with post-Columbian European and African sources at different times, with few cases of prolonged isolation. With our results we emphasize the importance of including under-studied regions of the continent in high-resolution genetic studies, and we illustrate the potential of SNP chip arrays for informative regional-scale analysis.

Introduction

The genetic diversity of the Americas has long been underestimated due to the paucity of population samples analyzed with high-resolution markers. Over the past two decades, population studies have focused on uniparental markers, predominantly typed at low resolution (reviewed in Bisso-Machado et al. 2012). Recent studies are increasing the coverage of the continent with high-resolution genomic data from ancient remains and living populations. The results confirm previous findings at a continental scale, such as a post-Last Glacial Maximum entry of a small founding population, a major migration ancestral to all living Native American groups from North to South America (Tamm et al. 2007; Reich et al. 2012; Achilli et al. 2013; Raghavan et al. 2015; Llamas et al. 2016; de la Fuente et al. 2018), and further layers of population structure and admixture as suggested by the analysis of ancient DNA. These demographic dynamics include an early diverging branch reconstructed for ancient North American sites (Scheib et al. 2018), which did not reach South America (Posth et al. 2018), and an enigmatic signal of Australasian ancestry recovered only in some populations of South America (Skoglund et al. 2015; Moreno-Mayar et al. 2018). The early population differentiation experienced after the initial entry into the continent resulted in different ancestries, such as the “Mixe” (Moreno-Mayar et al. 2018) or the

"ancient Californian Channel Islands" (Posth et al. 2018), as reconstructed by admixture graph methods. It is difficult to trace how these ancestral genetic components have survived in living populations, as there is a lack of dense sampling of populations with a high proportion of Native American ancestry. This also impacts our understanding of pre-colonial dynamics at a local scale, with only a few studies reporting good sampling coverage for targeted regions (Arias, Barbieri, et al. 2018; Harris et al. 2018).

In South America, genetic studies have robustly recovered a substantial differentiation between the Andes and Amazonia, which has been framed within a model of large communities connected by gene-flow in the Andes vs. small, isolated communities in Amazonia (Tarazona-Santos et al. 2001; Fuselli et al. 2003; Barbieri et al. 2014). This model builds on evidence for major complex societies in the Andes (culminating with the well-known but short-lived Inca Empire) which fostered population movements and connections, counterbalanced by the traditional view of the Amazon Basin as the homeland of small, isolated tribes. The latter view is challenged by increasing evidence of large-scale societies (Heckenberger and Neves 2009; de Souza et al. 2018), the role of rivers as primary routes for gene-flow (Arias, Barbieri, et al. 2018), and the presence of important centers of plant domestication (Clement et al. 2010). To gain a better representation of the highly diverse cultural landscape of Amazonia, a more intense archaeological effort is needed, together with a more fine-grained sampling of living and ancient human populations.

In particular, this model of South American genetic structure typically overlooks the Pacific Coast, a key context for the early migration history of the continent (Dixon 2013) and the cradle of the earliest complex societies in South America, from 3000 BCE (Stanish 2001). Recent studies have begun to investigate human variation on the Pacific Coast through ancient DNA (Fehren-Schmitz et al. 2010; Fehren-Schmitz et al. 2014; Valverde et al. 2016) and by sampling urban areas (Sandoval et al. 2013; Cabana et al. 2015; Harris et al. 2018). To fill out this picture, however, requires further, complementary genetic studies on living populations, especially from non-urban areas.

Language diversity is a further variable used to identify population groups, and which can correlate with population relatedness. The diffusion of major language families has been associated with demographic movements by some scholars (Renfrew and Bellwood 2002; Diamond and Bellwood 2003). Genetic studies appear to have validated this association for some of the largest language families of the world (de Filippo et al. 2012; Lipson et al. 2014; Haak et al. 2015), but no strong candidates are found in South America. Previous genetic work (Sandoval et al. 2016; Barbieri et al. 2017) has evaluated alternative models of cultural vs. demographic diffusion for Quechua, the most spoken language family of the Andes, present also in small pockets in the Amazonian lowlands (Cerrón-Palomino 2003). These studies, based on uniparental markers, revealed intense contact routes within the southern highlands, but not in northern regions nor in neighboring Amazonia. Relatively few genetic studies have addressed the diffusion of the main language families of Amazonia (notably Arawak, Tupí or Carib), although very recent research does focus on sub-branches or smaller families (Arias, Barbieri, et al. 2018; Schroeder

et al. 2018). Some scholars suggest that the main driver in the diffusion of Arawak was cultural contact alone (Hornborg et al. 2005). The particularly fragmented distribution of the three major language families across much of lowland South America (Epps 2009) calls for more fine-grained sampling to test for potential connections between their speaker populations.

Here we focus on western South America with 176 new genetic samples from populations with different cultural, linguistic and historical backgrounds, to investigate environmental and cultural influences on population genetic structure over three ecogeographic domains: the Andes, Amazonia and the Pacific Coast. We first explored the dataset to understand effects of early migration and ancient structure in underrepresented regions of the continent. We then focused on more recent historical layers, with the following three goals: 1) Evaluate the genetic impact of major complex societies, which arose in two main focal regions: the North Coast of Peru, and the Andean highlands of Central and Southern Peru and northern Bolivia. Large-scale societies possibly left a trace in the demographic profiles of indigenous populations, associated with high population density (Goldberg et al. 2016), but the extent of long-distance population movements and the origins of the populations that developed such societies remain largely unknown. 2) Describe the diffusion mechanisms of major language families. We aim at tracing genetic connections over the scattered and widespread diffusion of representative Andean and Amazonian languages, focusing in particular on a vast region where different varieties of Quechua are spoken. 3) Reconstruct the demographic events over the last five centuries since European contact, and how they impacted upon different South American populations. Gene flow from European and African sources can be easily distinguished in the genomic ancestry of the American populations (Gravel et al. 2013; Chacón-Duque et al. 2018). The timing and intensity of the European-mediated admixture has been estimated for urban, heavily admixed regions (Homburger et al. 2015; Harris et al. 2018), but has yet to be investigated systematically across South America.

Results

Continental-scale population structure

The newly genotyped data from western South America and Mexico were merged with available Native American population samples similarly genotyped with the Human Origins Affymetrix SNP chip (Fig. S1). Spanish, Italian_North and Yoruba population samples were included to distinguish admixture from European and African sources (Table S1). The comparative dataset includes 426 individuals and a total of 597,569 SNPs. We first investigated continental ancestry structure by means of ADMIXTURE analysis, with a dataset pruned for Linkage Disequilibrium (LD) of 232,755 SNPs. The cross-validation error was lowest for K=3 (Fig. S2), indicating that the clearest ancestry signal is the one that separates African, European and a shared Native American ancestry. Further levels of K were considered to explore structure within the Native American component. At K=4 a new component is found in most

Amazonian populations, while at K=5 the Xavante are distinguished from the other populations (consistent with their high levels of genetic drift, as described by the diversity values discussed later). At K=6 the Amazonian populations divide further into one component common to the Kichwa from northern Ecuador (Kichwa Orellana) and neighboring Colombian populations from the eastern slopes of the northern Andes (the light green ancestry component in Fig. S2, designated "Amazonia North") distinct from a further component common to the remaining Amazonian populations (dark green ancestry component, designated "Amazonia Core"). At K=7 the North American populations are distinguished by a separate component (purple color). At K=8 a further ancestry component is distinguished in the Central-Southern Andes (dark blue). At further levels of K the cross-validation error begins to increase appreciably.

This structure is reproduced in the principal component analysis (PCA, Fig. 2), performed with a set of 2,545 SNPs ascertained for Karitiana in the initial Human Origins assembly — Panel 7 in Patterson et al. (2012). The first dimension separates samples from both "Amazonia North" and "Amazonia Core" from the rest of the Americas. The second dimension separates off "Amazonia Core", the third separates off the North American samples, and the fourth dimension separates off the Central-Southern Andes. Individuals from different locations on the North Coast display a wide range of variation and in all dimensions partially overlap with the North-East Andes of Peru.

Broad population relationships can be estimated by the F_{ST} distances between populations, here visualized with a Multi Dimensional Scaling (MDS) on three dimensions and a heatplot (Fig. S3). The population structure corresponds to a broad division between the following macro-regions: North America, Pacific Coast + Andes, and Amazonia, the last of which can be further divided between the proposed "Amazonia North" and "Amazonia Core" components.

To explore the relationship between the major components retrieved in western South America, we performed coalescent simulations. We modeled a scenario of population split with broad-time split priors, with and without migration (Fig. S4). We chose three relatively unadmixed populations with a minimum sample size of 10, representative of the three major ancestry components: Coast (Sechura_Tallan), "Amazonia North" (Kichwa Orellana), "Amazonia Core" (Wayku). We included a second population from the "Amazonia Core" (Karitiana) to reproduce the ascertainment bias of the SNP set from Panel 7 in Patterson et al. (2012). The posterior ABC analysis supports a model with migration vs. a model without migration (in 92.2% of the cases), with a first split of Coast, "Amazonia North" and "Amazonia Core" at 6-7 kya, a second split between the Coast and "Amazonia North" broadly inferred between 4 and 7 kya, and a third split between Wayku ("Amazonia Core") and another population of the "Amazonia Core" (Karitiana) at 2-2.5 kya. Migration rates (the proportion of migrants per generation) between the three main groups vary between 0.01 (Kichwa Orellana and Wayku), and 0.018-0.2 (Kichwa Orellana and Sechura_Tallan, Wayku and Sechura_Tallan).

Demographic reconstructions and drift

To assess whether we can distinguish different demographic trajectories for the populations considered, we analyzed runs of homozygosity (ROH) blocks. ROH blocks are inherited from a common ancestor, and their length is inversely proportional to the number of generations since the split. ROH blocks that result from a recent bottleneck will tend to be longer, as there are fewer recombination events. ROH blocks are also informative about effective population size (N_e), as populations with low N_e tend to have more extended ROH than those with high N_e (Kirin et al. 2010). Here ROH blocks were divided either into two length classes as suggested by Pemberton et al. (2012) (Fig. 3A), or six bins as in a previous study of Native American populations (Schroeder et al. 2018) (Fig. 3B, S5). All of the populations have an excess of short ROH (<1.6 Mb); this excess was lower in those populations exhibiting more European admixture (Fig. 3A). The short ROH likely reflect the strong bottleneck experienced by the founding population of the Americas, as previously noted (Pemberton et al. 2012; Mooney et al. 2018; Schroeder et al. 2018). The long ROH classes are differently distributed among populations, regardless of their geographic proximity or, more broadly, their ecogeographic domain. The populations with the highest proportion of large ROH are the Karitiana, Xavante, Cabécar and Pima, as previously noted (Pemberton et al. 2013; Ceballos et al. 2018); here they are grouped in the "Published data" panel of fig. 3B. We can further distinguish populations with fewer ROH blocks longer than 2-4 Mb (group 1 in Fig. S5). Some of these populations have more European/African admixture (group 1a), but other populations are considerably less admixed with Europeans (groups 1b and 1c in particular). The absence of long ROH implies that these populations did not share a recent bottleneck: in this group are populations from Amazonia (Cocama), most of those from the Coast, some from the Andes (La Jalca, Cusco2, Parán, Puno) and the Yaquis from Mexico. Admixture between native populations could have impacted the distribution of this class of ROH across domains. Populations with a peak of ROH length at 4-8 Mb may have experienced a recent bottleneck (group 2): this is common in Amazonia (Kichwa Orellana from Ecuador, Cofán in Colombia, LoretoMix in Peru). N_e was also estimated through our simulations for the three populations taken as representative of the Coast (Sechura_Tallan), "Amazonia Core" (Wayku), and "Amazonia North" (Kichwa Orellana). The N_e of Sechura_Tallan (estimated at ~2500-3000) is larger than the one for Kichwa Orellana and Wayku (1000-2000, Fig. S4B).

Population internal diversity and drift were also evaluated by calculating estimates of consanguinity per individual (coefficient F), visualized in Fig. S6A. Published data for the Karitiana, Xavante and Cabécar have the highest levels of consanguinity. Overall, consanguinity is slightly higher in Amazonian populations than on the Coast. The level of consanguinity is directly correlated with the proportion of Native American ancestry estimated by supervised ADMIXTURE analysis (Pearson's coefficient = 0.94): Fig. S6B displays this correlation, with (as expected) slightly lower values of F by proportion of Native American ancestry on the Coast and in the Andes, and slightly higher F values in Amazonia,

with a few individuals from the Inga, Yaquis, and Cusco2 populations who have less Native American ancestry but high levels of consanguinity.

Recent contact from haplotype sharing networks mirrors linguistic connections

To investigate recent historical layers of contact we analyzed identity by descent (IBD) segments. Identical blocks between individuals correspond to shared ancestry, with longer blocks corresponding to recent shared ancestry. Fragments shorter than 5 cM are shared between almost all pairs of Native American populations (data not shown), in agreement with other studies of South American populations (Harris et al. 2018). This diffused pattern of sharing might reflect the reduced genetic diversity of the continent from the initial founding bottleneck (resulting in a high overall level of consanguinity, see Palamara et al. 2012; Mooney et al. 2018). To focus on the most relevant sharing patterns, a threshold of 5 cM was applied, and population pairs which shared only one fragment were not considered (to reduce chance effects).

Fig. 4A shows the overall pairwise sharing patterns, while Fig. 4B includes only those pairs for which the number of shared blocks (adjusted for population size) is higher than the median, to further highlight the most significant sharing networks. The highest impact of sharing events can be found along the diagonal in Fig. 4A, within the various regions covered: Amazonia, North Coast, North-East Andes and Central-Southern Andes. The number of sharing events between pairs of populations is inversely proportional to their geographic distances (Fig. S7, Mantel Test with Spearman correlation = -0.62, $p<0.01$), as expected, but there is a high degree of sharing between some geographically distant pairs. For example, sampling locations along the Coast, where the total length of shared blocks is greater, span a longitudinal distance of almost 700 km, while sampling locations in Central-Southern Andes, where the total length of shared blocks is lower, cover a total distance of ~1000 km. We find a significant connection between "Amazonia Core" populations, which share high numbers of large blocks over a long distance (Fig. 4B). This sharing involves speakers of Cocama (a Tupí language) in Colombia, who share 30 IBD blocks longer than 10 cM with individuals from the "LoretoMix" group in Peruvian Amazonia, whereas the LoretoMix shares only 10 of such large IBD blocks with the neighboring Wayku. The LoretoMix includes three Cocama speakers, and only these three individuals share IBD blocks with the Cocama from Colombia despite a distance of more than 500 km separating the two sampling locations. These samples are marked with a red asterisk in the PCA (Fig. 2), where they are also close to the Cocama from Colombia. The strongest signal of relatedness is found between the neighboring Inga and Kamentsa populations from Colombia, who share numerous, long IBD blocks.

In North America, Yaquis share many long blocks with Pima (both speak languages from the Uto-Atzecan family), at a distance of 250 km. Finally, numerous shorter fragments are found to be shared between Amazonia and the Andes, in particular between speakers of languages within the Quechua family: Kichwa Orellana and Wayku are connected with populations of the North-East and Central-

Southern Andes. This signal of particular interest as it shows how recent contact in the form of IBD sharing can be used to reconstruct the signature of language expansions in some regions of the Americas.

Post-Columbian admixture with Europe and Africa

We examined the uniparental data (in terms of haplogroup frequencies) for a first overview of the proportion of Native vs. non-Native American ancestry in each population (Table S2). The typical Native American haplogroups for mtDNA are A2, B2, C1 and D1 (plus the less frequent D4h3 and X2a, the latter not present in our dataset), while for the Y chromosome they are Q and C3 (Bisso-Machado et al. 2012). Fig. S8 shows that in most groups the frequency of Native American mtDNA haplogroups is 100%; the exceptions are groups from the Coast (Cao and Tumbes), which have a few individuals assigned to the African haplogroups L3 and L2 (Table S2, marked as "others" in Fig. S8). The frequency of Native American haplogroups is overall lower in the Y chromosome than in the mitochondria, but it reaches 100% in all individuals in Amazonia Core, in the Central-Southern Andes and in some populations of the Coast. Non-Native American haplogroups (mostly R, of European origin, but also E, potentially of African origin) predominate only in Chulucanas, Tumbes, Cao, and La Jalca (Table S2, marked as "others" in Fig. S8).

A supervised ADMIXTURE analysis was then performed to investigate the proportion of Native American ancestry per individual (Fig. S9). This analysis shows that several populations from all three ecogeographic domains display negligible proportions of European or African ancestry, confirming the results from the uniparental data. The proportion of Native American ancestry in the autosomal data, averaged per population, is roughly proportional to the average of female (Pearson's correlation = 0.34) and male Native American ancestry (0.67), with the latter being lower than expected in the admixed populations of the Coast in particular (Fig. S10). The proportion of European ancestry is uniformly distributed among individuals only in the North-East Andes populations. In all other populations that show evidence of European and African ancestry, the proportion of those ancestries varies widely at the individual level: this clearly suggests additional and more recent episodes of gene flow into these groups. For subsequent analyses of admixture (which are more robust for large sample sizes), populations were grouped according to similar Native American ancestry profiles, and outlier individuals were excluded (i.e. a single individual showing exceptional non-Native American ancestry among unadmixed individuals of the same population, as was the case in Sechura, Kamentsa and Cofán, and as indicated in the third column of Table S2). Furthermore, because the Colombian Inga clearly show structure with respect to their ancestry (Fig. 1), we separated the highly admixed individuals into an Inga_Admixed population, and merged the less admixed Inga individuals with the neighboring Kamentsa.

We used an f_3 admixture statistic of the type f_3 (Target; Source1, Source2) to confirm admixture events between Native American populations and European and African sources, where the target population is a South American population for which the ADMIXTURE results suggest European or African

ancestry components. Source1 is a non-admixed Native American population (Xavante, Sechura_Tallan or Puno) and Source2 is either a European (i.e. Spanish) or an African population (i.e. Yoruba). Negative values of f_3 confirm the signal of European admixture for a few populations of the Coast, for the Mexican Yaquis, for all North-East Andes, for Cusco2, and for the Inga_Admixed (Fig. S11). African admixture appears in a subset of these populations, with the strongest signal in the Coast and in Inga_Admixed.

To further investigate the signal of recent admixture suggested by ADMIXTURE, we analyzed IBD blocks shared with Yoruba and Spanish sources. Sharing is detected in all three ecogeographic domains (Fig. S12). The largest number of blocks from the European source is found in Cao, Chulucanas, Tumbes, Cusco2, Yaquis, Inga_Admixed, Luya and UtcubambaSouth. The pattern from the IBD sharing agrees with the profile from the supervised ADMIXTURE, with some exceptions: in Kichwa Orellana and Wayku, the IBD blocks imply more European ancestry than the ADMIXTURE results do.

To explore the intensity and timing of post-European contact in our selection of populations we employed two methods, which date admixture based on different aspects of the data: MALDER (Loh et al. 2013) and wavelet transform analysis (WT, Pugach et al. 2011). Both methods are applicable to admixture events involving more than two source populations. We again used Yoruba and Spanish as proxies for the African and European source populations, respectively. The results are summarized in Fig. 5. Local ancestry along individual chromosomes was inferred using the RFMix method (Maples et al. 2013). With MALDER we ran the analysis to infer dates for both European and African admixture for all populations, regardless of the admixture proportions. With the WT-based method, meanwhile, for African admixture we inferred dates only if the proportion of African ancestry in a given population was over 1% (estimated based on RFMix). The dates inferred by WT are on average 8.7 generations earlier than those inferred by MALDER (SD = 4). It has been shown previously (Pugach et al. 2018) that this discrepancy is expected with continuous admixture or multiple pulses of gene flow from the same source, as MALDER is more sensitive to recent admixture events.

The dates inferred for European admixture are in most cases more recent than those for African admixture, reflecting an admixture history protracted through time for the European ancestry source. The most recent dates (for both African and European admixture) correspond to 7-8 generations ago for MALDER and 8-10 generations ago for the WT method. The older dates are found in Amazonia, in particular in our “Amazonia North” (Kichwa Orellana and Inga) and in Wayku, where the admixture is estimated to have happened between 1650 and 1700. Here the dates from MALDER and WT mostly overlap.

Discussion

We generated genome-wide data with the Affymetrix Human Origins array for 176 individuals from 25 populations of North America and western South America, and analyzed these data together with published data from representative populations of the continent. Our strategy in collecting and analyzing the data can be summarized under three major objectives. 1) To investigate patterns of genetic diversity within and between the three main ecogeographical domains of western South America (the Andes, Amazonia and the Coast), especially in understudied regions and in transitional environments. 2) To retrace past connections between and within the domains, and to evaluate to what extent the genetic landscape of South America was impacted by the last and largest complex societies of the pre-Columbian period, and is reflected in the distributions of indigenous language families. 3) To reconstruct the timing of admixture events from European and African sources *after* Columbus, and to identify differences in the chronology of such admixture within each of the three main domains.

For the first objective, we investigated Native American ancestry at a continental scale. One major Native American ancestry component is shared by all populations, as seen in the ADMIXTURE plot (Fig. S2, K=3, associated with the lowest cross-validation error), in line with results from other living populations and from ancient DNA, which support an early entry as a single major migration (Raghavan et al. 2015; Llamas et al. 2016; Harris et al. 2018; Moreno-Mayar et al. 2018). This is not unexpected, as further multiple migration effects are reflected in more subtle genetic signals. The diversification of further ancestry blocks from the initial single Native American gene pool does not bear traces of a north-south gradient of differentiation, or of serial founder effects (Fig. S3). A previously observed early diverging component similar to the Mixe (Moreno-Mayar et al. 2018) or the ancient Californian Channel Islands (Posth et al. 2018) is not captured by our data, which focuses more on genetic diversity within South America. Amazonian ancestry is further split into two components: one more widespread in the Amazon Basin (here called "Amazonia Core"), the other in the piedmont populations of Ecuador and south-western Colombia ("Amazonia North", Fig. 1 and 2). This latter component is strongly differentiated: simulation analyses suggest that the "Amazonian North" component split from the Coast and the "Amazonia Core" at an early diverging stage, at least 4,000 years ago (Fig. S4). Even the potential drift associated with the small N_e cannot entirely account for the divergence between these populations. This "Amazonian North" ancestry is found in a transitional environment: this region spanning parts of Colombia and Ecuador is in fact geographically close to the Northern Andes, but its populations are traditionally associated with the Amazonian cultural domain. An early human settlement of Ecuador and northern Peru (between 16.0 and 14.6 kya) has previously been inferred from high-resolution mtDNA data (Brandini et al. 2018), in line with the archaeological record (Dillehay et al. 2017). Meanwhile, the presence of pockets of diversity in Ecuador and Colombia is paralleled by the presence of distinctive Native American lineages, such as Y chromosome haplogroup C3, otherwise rare in the continent (Mezzavilla et al. 2015). This haplogroup is also reported for other populations in

Colombia (Arias, Schroeder, et al. 2018) and is found in the sample from Ecuador (Fig. S8, Table S2), but with the available data we cannot confirm that it corresponds to the sublineage C3 of haplogroup C.

Finally, populations from the Coast and the Central Andes (both north and south) show close genetic proximity to each other, as visualized by the PCA in Fig. 2 and by sharing the same ancestry component profile up until K=6 in Fig. S2. This suggests a common origin and/or extensive contact, which may be associated with a coastal migration route and a colonization process from the coast inland into the highlands (Wang et al. 2007; Rothhammer and Dillehay 2009; Reich et al. 2012; Rademaker et al. 2014; Harris et al. 2018). Previous analyses have already noted the common history of these two regions, possibly dating to an early settlement ~12,000 years ago (Harris et al. 2018).

For our second objective, on connections within and between domains, we explored signatures of demographic history and haplotype sharing patterns. The ROH variation profile of most populations from the Coast and the Andes is consistent with a history of a relatively large population size, with some exceptions (Sechura, Narihuala, Cusco) that may have experienced isolation and drift only very recently (Fig. 3, Fig. S5). The long-term presence of large-scale state societies in the Andes and on the Coast can be expected to have promoted gene flow across wider geographical scales and merged previously structured populations, contributing to the higher genetic diversity of the current inhabitants. On the North Coast of Peru, the Moche culture was one of the largest entities from the 1st century CE, followed by the Chimú from the 12th century (Quilter 2013). Their political influence over the coast would have overcome the stretches of desert that separate the main river valleys, and the Humboldt current and wind regime that make long-distance seaborne trade difficult. In the Chachapoyas region (North-East Andes), a number of structured societies flourished from the 12th to the 15th centuries (Church and Von Hagen 2008). In the Central-Southern Andes, the Wari and Tiahuanaco 'Middle Horizon' (c. 500-1000 CE) and especially the Inca 'Late Horizon' (c. 1470-1532 CE) established vast networks that mobilized and moved large labor forces for agricultural production (terracing, irrigation, raised fields), operated resource exchanges through camelid caravans, and resettled populations as explicit state policy (Isbell 2008; Quilter 2013; D'Altroy 2014). The impact of the Wari and Inca Empires is widely associated also with the diffusion of the two main surviving Andean language families, Quechua and Aymara (Adelaar and Muysken 2004; Heggarty 2008).

The Coast and our two Andean sub-regions share a similar ancestry (as discussed above) and a similar history of large population size, but they are differentiated at a finer scale, with localized patterns of IBD segment sharing. By contrast, the Amazonian populations in most cases have longer ROH blocks and overall high levels of consanguinity. This could reflect the model first proposed by Tarazona-Santos et al. (2001): larger groups in the Andes vs. small, isolated groups in Amazonia. Nevertheless, by including more populations from a wider range of cultural and geographical backgrounds, we find exceptions to this model, with some Amazonian populations characterized by a smaller number of larger blocks, belonging to group 1c and group 2 in Fig. S5. The populations of Amazonia therefore display

different demographic histories, rather than a uniform history of small sample size (according to the ROH profiles), and are connected by sharing of IBD blocks within the region. Moreover, Amazonian populations also show long-distance sharing of large and short fragments with the Andes and the Coast (Fig. 4), which is not consistent with the traditional portrait of isolation between Amazonian populations. This genetic diversity complements the evidence from other disciplines that the region was also home to dynamic, non-isolated population groups (Arias, Barbieri, et al. 2018). In particular, the linguistic diversity of Amazonia includes not just language isolates but major, expansive language families, with far-reaching geographic distributions (Epps 2009). There is also linguistic evidence for intensive interactions in convergence zones, and (more weakly) across Amazonia as a whole (Dixon and Aikhenvald 1999).

We explored these potential connections by checking for gene-flow among speakers of the same language or language family. An interesting case is represented by the speakers of Cocama, a language of the Tupí family. The ROH profile for the Cocama of Colombia is lacking in long ROH segments (Fig. 3, Fig. S5), suggesting no recent bottlenecks or isolation. The analysis of shared IBD segments reveals a long-distance connection between this population and geographically-distant populations of Peruvian Amazonia (Fig. 4). In particular, three Cocama speakers included in the LoretoMix sample from Peru are genetically close to the Cocama of Colombia (Fig. 2). Archaeological and ethnohistorical evidence indicates that the ancestors of the Cocama and Omagua were widely dispersed in pre-Columbian times, inhabiting large stretches of the Amazon Basin and several of its upper tributaries (Lathrap 1970; Michael 2014). Thus, the sharing of IBD segments as well as the lack of long ROH in the Cocama could be explained by large, widespread populations that were connected in pre-Columbian times. Alternatively, more recent migrations could have carried the Cocama language between Colombia and Peru. Both time-frames and both scenarios suggest a parallel between genetic and linguistic history, with language acting as a preferential tracer of population mobility.

Weak evidence for long-distance linguistic connections is observed not only within Amazonia, but also between Amazonia and the Andes. This is the case for Quechua-speakers of lowland Ecuador (Kichwa Orellana) and lowland north-eastern Peru (Wayku), who share relatively short IBD fragments with the Central-Southern Andes and North-East Andes respectively. Previous results based on Y chromosome haplotype sharing did find a similar pattern of connections between lowland Quechua-speakers in Ecuador and north-eastern Peru, but did not find such long-distance connections with the Central and Southern Andes (Sandoval et al. 2016; Barbieri et al. 2017). These different results can possibly be justified by sex-biased gene-flow (i.e. less male mobility), which should be further investigated with denser sampling and high-resolution mtDNA genome sequences. Overall, this new genomic evidence points towards a demographic connection behind the diffusion of Quechua varieties not only in the southern highlands, as previously attested (Barbieri et al. 2017), but also in the north, across ecogeographic domains.

Finally, for the third focus we explored the traces of post-colonial history and the impact of European mediated gene-flow (from Europe and from Africa through the slave trade) in the different ecogeographic regions. In our newly reported samples we find a high proportion of Native American ancestry, with some populations showing no detectable post-Columbian admixture in all three ecogeographic domains (Fig. 1) and a high proportion of Native American mtDNA and Y chromosome haplogroups (Figs S8 and S10). These results are in agreement with previous studies on ancestry proportions among Peruvian populations (Sandoval et al. 2013; Harris et al. 2018). A high Native American ancestry proportion is even observed for the Coast, even though the traditional fishing/trading economy (Sandweiss 2008) might have been expected to introduce gene flow also from non-Native American sources. Importantly, our sampling strategy was guided to avoid individuals who self-reported any grandparent or parent of European, African or Asian descent, thereby introducing a first filter for recent admixture. Nevertheless, this strong Native American ancestry reveals the potential of undersampled regions of the Americas for further exploring pre-Columbian genetic history.

We used two different methods to estimate the date of admixture with European and African sources (Fig. 5). While simulations show that in simple one-pulse admixture situations both MALDER and WT-based methods perform equally well for both recent and older admixture times, the dates inferred by both methods are not concurrent in more complex admixture scenarios, involving either multiple pulses or continuous gene flow (Pugach et al. 2018). MALDER is more sensitive to the most recent admixture event experienced by a population, while the WT method is more sensitive to older admixture events, and tends to give intermediate dates when there are multiple admixture pulses (Pugach et al. 2018). Here, the WT method consistently returned older dates than MALDER, suggesting multiple and/or continuous admixture. The oldest WT dates may reflect the initial episode of admixture experienced by some populations during the earliest colonization by the conquistadors, historically dating to the mid-16th century. Of these populations, the majority have much more recent MALDER dates of 7-8 generations ago (around the end of the 18th century), i.e. the populations of the Coast, the admixed samples in the highlands from Cusco (Cusco2), and the Yaquis of Mexico. It is reasonable to assume that the contact with Europeans began earlier in these regions: the recent admixture dates may be describing either continuous admixture or a second, more recent pulse of admixture (not necessary from Europeans, but also from local *mestizos*). This would be compatible with the admixture profile of Peru as reconstructed by a recent study, where the major pulse of European admixture occurred during the 19th century, after the impact of the war of independence in Peru (Chacón-Duque et al. 2018; Harris et al. 2018). Nevertheless, not all populations fit this profile of a recent admixture pulse: in "Amazonia North" and in North-East Andes (where La Jalca is the most isolated location), MALDER recovers older admixture dates, between 15 and 11 generations ago, which often overlap with the WT dates (Fig. 5). These potential pockets of isolation from further pulses of admixture, which lasted for three centuries, indicate different patterns of integration, or a less continuous gene flow from individuals who carry

European ancestry. The admixture dates around 1650-1700 are in agreement with historical records of early intrusions of Europeans (including missionaries) into Peruvian and Ecuadorian lowland rainforests (Sandoval et al. 2016).

Finally, studies of ancient DNA have shown that as much as one third of the ancestry in modern Native Americans could be traced to western Eurasia (Raghavan et al. 2014). Similarly, modern-day Europeans were found to be a mixture of three ancestral populations, one of which was a population deeply related to Native Americans (Lazaridis et al. 2014). These findings imply that European (or more accurately, Eurasian) ancestry found in modern-day Native Americans may not have been acquired exclusively through admixture during the post-Columbian period, but instead may reflect a much deeper origin. It is therefore possible that the WT method is picking up this signal of shared ancestry, which predates European colonization, and hence infers dates for some populations that are too early to be consistent with the first appearance of the conquistadors in the Americas, only after 1492.

Admixture with African sources appears with relatively older dates and shorter fragments (Fig. S12), as it did not continue through time with the same intensity as the admixture with European sources (mostly through *mestizos*). It is also possible that the African component was incorporated principally through European-mediated gene flow, as individuals in our samples who carry African ancestry always carry European ancestry as well (Fig. S9). These cases indicate some degree of isolation over the last two centuries from the admixture that occurred during the periods of Spanish colonial rule (from 1532 to 1821) and of slavery (which largely overlapped), and replicate historical records for African slavery in Peru (Arrelucea Barrantes et al. 2015). The proportion of African individuals in the population was at its peak before 1800, but declined rapidly in proportional terms during the nineteenth century. In the colonial period and indeed thereafter, the African population was concentrated on the coast, where it was exploited for plantation agriculture. Finally, the incorporation of the African genetic component was typically mediated by European males, while during the period of slavery marriage between people of African descent was hindered by the Spanish colonial regime. The *Sistema de Castas* enforced by that regime segregated both Africans in plantations and indigenous settlements from European and *mestizo* groups, at least until the early and mid-colonial periods (Socolow 2015).

In conclusion, by targeting key regions of western South America and focusing on high-resolution SNP array data, we are able to reveal demographic histories, ancient structure and recent connections between different ecogeographic domains. These connections are particularly interesting for Amazonia, traditionally portrayed by genetic models as a region of small isolated communities.

We also note how certain population samples widely analyzed in recent genetics literature, e.g. the Karitiana and Xavante, exhibit high levels of genetic drift in comparison to our newly generated dataset — see the analyses of population relationship (Fig. S3B) and of within population diversity (Fig. 3, Fig. S6). It is important to stress that inferences on Native American prehistory should not be drawn

exclusively from such divergent populations with many closely-related individuals, but should instead include more diverse populations from different regions and different cultural and demographic backgrounds, in order to capture the diversity of the continent (Homburger et al. 2015; Bolnick et al. 2016).

Materials and Methods

Sample collection

Samples were collected during anthropological fieldwork expeditions by R.B. and C.Z. (Ecuador, 2007), L.A. (Colombia, 2012), C.B., R.F., J.R.S., and O.A. (Peru, 1998, 2007, 2009, 2014, 2015), and A.A.C. and R.S.O. (Mexico, 2016). The sampling collection and the project were approved by the Ethics Committee of the University of San Martín de Porres, Lima (Comité Institucional de Ética en Investigación de la Universidad de San Martín de Porres — Clínica Cada Mujer, Oficio No. 579-2015-CIEI-USMP-CCM, 12/05/2015), the ethics committee of the Universidad del Valle in Cali, Colombia (Acta No. 021-010), the Ethics Commission of the University of Leipzig Medical Faculty (232/16-ek), the Ethics Committee of the University of Jena (Ethik-Kommission des Universitätsklinikums Jena, Bearbeitungs-Nr. 4840-06/16), the Research Council for Science and Technology (Consejo Nacional de Ciencia y Tecnología - CONACyT, grant # 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán Ref.: 1518), and the National Commission for Scientific Research of the Mexican Institute for Social Security (IMSS; CNIC Salud 2013-01-201471). All methods were performed in compliance with the rules of the Declaration of Helsinki. The samples analyzed in this study represent only a small fraction of the population living in the target regions of Mexico, Peru, Colombia and Ecuador, and so is only partially representative of the complex demographic history of these regions and of their inhabitants' ancestors.

Details of the sampling collection and DNA processing are reported in Arias, Barbieri, et al. (2018) for the four Colombian population samples and in Barbieri et al. (2017) for the Peruvian samples from Luya, La Jalca, Huancas, UtcubambaSouth (department of Amazonas) and Wayku (department of San Martín). The samples identified as "Cusco2" correspond to individuals who were sampled in the urban districts of San Sebastián and San Jerónimo (Cusco, Peru); details of the sampling are described in Sandoval et al. (2018). Samples identified as Ecuador Kichwa were previously analyzed in search for a genetic variant associated with lipid metabolism (Acuña-Alonzo et al. 2010). The other population samples have not been previously reported or described.

Samples from the population named "LoretoMix" include three speakers of Cocama (a language of the Tupí family), one of Chamicuro (Arawak), one of Shawi (Cahuapanan) and two of Muniche (a language isolate). These samples were collected in various locations within the department of Loreto in the

Amazonian region of north-eastern Peru, and merged into one population after verifying their genetic affinity. The population samples from Cusco and Cusco2 consist of speakers of southern Quechua. The population sample labelled Puno (department) is made up of five speakers of southern Quechua and two of Aymara, collected on the islands of Lake Titicaca and merged into one population after verifying their genetic affinity. Parán is a community located in the highlands of the department of Lima, who speak Spanish. The population samples from the North Coast of Peru include participants from rural areas and fishing communities who speak Spanish. The various population samples have been identified by the names of the towns or provinces where the samples were collected. Samples from the population named Kichwa Orellana include individuals sampled from the rural parish of San José de Guayusa, in the province of Orellana, in the Amazonian lowlands of Ecuador. The community speaks a variety of lowland Kichwa (the local name of Quechua), and includes individuals who recall relationships with Shuar communities from southern Ecuador. Samples from Yaquis were collected in the state of Sonora in north-western Mexico in the community of Tórim. People living there continue the culture and traditions of the Yaqui Nation and speak Yaqui, a language of the Uto-Aztecan family. The Mexican sample was included as a comparative source of genetic diversity from indigenous North America. For our population samples we associated a linguistic affiliation accounting for the majority of the community members: this was documented during the anthropological fieldwork, cross-checked by fieldwork assistants, and reviewed by P.H. for accurate historical linguistic contextualization.

The samples have been subdivided into seven groups by country and macro-region (Mexico, Colombia Amazonia, Ecuador Amazonia, Peru Amazonia, Peru North-East Andes, Peru Central-Southern Andes, Peru North Coast – Table S1). Individual information with details on the population, language spoken and geographic grouping is listed in Table S2. The sample locations for each population are shown in Fig. 1 and in more detail in Fig. S1.

Data generation and screening

The DNA samples were screened and quantified with a Nanodrop spectrophotometer and Qubit fluorometer, and visually assessed by gel electrophoresis at the laboratory of the Department of Archaeogenetics of the Max Planck Institute for the Science of Human History in Jena. Sample genotyping was performed by ATLAS Biolab in Berlin on the Affymetrix Axiom Human Origins array (Patterson et al. 2012). Genotyping data were processed using Affymetrix Genotyping Console v4.2.0.26. In total 188 samples were genotyped and genotyping call rates were >98.5% for all SNPs. The final dataset comprised 633994 SNPs. PLINK v1.90b5.2 (Chang et al. 2015) was used to calculate missing genotype rate with the command --missing. Average missing calls per sample is 0.005, with a maximum of 0.023 (see Table S2). After the merging with the comparative dataset, average missing call is 0.002 with a maximum of 0.02. PLINK was then used to calculate the inbreeding coefficient F (i.e. (<observed hom. count> - <expected count>) / (<total observations> - <expected count>)) and Pi_Hat values (degree of relatedness as Proportion of IBD, i.e. P(IBD=2) + 0.5*P(IBD=1)) between pairs of

individuals, filtering for minimum allele frequencies of 0.05. One individual with a high F value was excluded and only one individual was kept in eight pairs with Pi_Hat > 0.5. The same sample was included twice for cross-reference (CH008): we found 700 different nucleotide calls between the two, which correspond to an error rate of 0.1% in the genotyping. One duplicated sample was found, probably due to mislabeling. The final screened dataset consists of 176 individuals which were included in the analysis. See Table S2 for the details of the individuals filtered out.

Data availability

To access the genotyped data, researchers should send a signed letter to C.B. containing the following text: ‘‘(a) I will not distribute the data outside my collaboration; (b) I will not post the data publicly; (c) I will make no attempt to connect the genetic data to personal identifiers for the samples; (d) I will use the data only for studies of population history; (e) I will not use the data for any selection studies; (f) I will not use the data for medical or disease-related analyses; (g) I will not use the data for commercial purposes.’’

Uniparental markers

Mitochondrial haplogroups were assigned with Haplogrep (Weissensteiner et al. 2016), limiting the call to major haplogroup nodes, given the uncertainty arising from the low number of mtDNA SNPs included in the Human Origins array. Y chromosome haplogroup assignment was performed with the yHaplo software (Poznik 2016). Data was cross-checked with available published mtDNA and Y chromosome data for the same individuals, assigned via direct genotyping/sequencing in previous studies (Barbieri et al. 2017; Arias, Barbieri, et al. 2018; Arias, Schroeder, et al. 2018): the SNPs available allowed the correct macro haplogroup to be detected in 97% of cases.

Merging

The newly generated dataset was merged with published Human Origins data from (Patterson et al. 2012; Lazaridis et al. 2014; Skoglund et al. 2015), selected to include populations representative of North and South America and of post-colonial African and European ancestry (Yoruba, Spanish and Italian_North were chosen for these analyses). Not all samples or populations were used for all analyses, as described for each analysis. Merging was performed with the mergeit command in AdmixTools (Patterson et al. 2012). A total of 597,569 SNPs were left after merging. New and published data locations are visualized on a map, which shows also the sample size for each population (Fig. S1).

Admixture analyses

We used the ADMIXTURE software (Alexander et al. 2009) to infer individual ancestry components and admixture proportions, after performing LD pruning with PLINK. The LD pruning included the following settings, which define window size, step and the r^2 threshold: –indep-pairwise 200 25 0.4

(Pugach et al. 2018), leaving 232,755 SNPs. We preferred a conservative approach with rather stringent setting parameters to robustly resolve the structure of the dataset.

We ran ADMIXTURE for values of K from 2 to 12, with 100 runs per K. Results of the ADMIXTURE runs are visualized with PONG (Behr et al. 2016). We used the cross-validation procedure implemented in ADMIXTURE to find the best value of K, and verified a regular, unimodal distribution of likelihood behind each K to exclude hidden multi modal results. The support for each K is indicated by the number of runs which return the same cluster composition. Population outliers such as Pima, Karitiana and Cabécar were excluded from this analysis — only Xavante was kept as a reference for Amazonian populations. Supervised ADMIXTURE (K=3) was performed to estimate the proportion of African, European and Native American ancestry per individual, keeping Yoruba, Spanish and Xavante (the latter known to be mostly unadmixed with European and African sources) as proxies for the parental groups.

We calculated f_3 statistics as a formal test for admixture, using the same European/African parent populations as suggested by the results of the ADMIXTURE analysis, and with three unadmixed Native American populations with sample size larger than 6 (Xavante for Amazonia, Puno for the Andes, Tallan together with Sechura for the Coast). The qp3Pop command from the AdmixTools package (Patterson et al. 2012) was used to run f_3. For each target population, the highest f_3 value was kept (corresponding to the best choice of Native American parental population among the three proposed).

PCA

Principal component analysis (PCA) was performed with smartpca of the Eigensoft package (Patterson et al. 2006). For this analysis we used a subset of SNPs ascertained in the Karitiana (Panel 7 as identified by Patterson et al. 2012), consisting of 2,545 SNPs that were heterozygous in a single Karitiana genome sequence. Smartpca was also used to calculate F_{ST} distances (Weir and Cockerham 1984) between populations, which were used to generate a heatplot of distances and a non-metric MDS (without outliers) in R with package MASS (Venables and Ripley 2002) and. We excluded outlier samples (Karitiana, Xavante, Cabécar and Pima) a posteriori, to investigate the overall continental structure.

Demographic simulations

We estimated the migration rates and separation times between Kichwa Orellana, Sechura_Tallan and Wayku using full genome coalescent simulations and then retaining a set of 2,635 informative sites ascertained as in the panel 7 of the Human Origin Affymetrix Chip. For these sites we calculated a set of D, f3 and F_{ST} statistics (Weir and Cockerham 1984, Patterson et al. 2012 - Table S3) using 10 individuals for 6 populations: Kichwa Orellana, Sechura_Tallan, Wayku, Karitiana, Chukchi and Yoruba. Simulations were run using the software scrm (Staab et al. 2015) according to the demographic scenario described in Fig. S4, following the priors reported in Table S4. To design the demographic scenario we started with an ancestral population (Anc) which splits between an African group (here represented by Yoruba) and an Out of Africa group (OoA). The OoA then splits between an Asian

outgroup (here represented by Chukchi) and a South American (SA) group. Priors were assigned for the following parameters: the N_e of each population, including the intermediate Anc, OoA, SA, KO-ST (ancestral to Kichwa Orellana and Sechura_Tallan), W-Ka (ancestral to Wayku and Karitiana); the split time for the Out of Africa (tOoA), the split between Asian and Americans (tSA0), the entry in South America (tSA, same broad priors as tSA0), the split between Wayku and Karitiana (tW-Ka), the split between Kichwa Orellana and Sechura_Tallan (tKO-ST, same broad priors as tW-Ka); and finally the migration rates between the three target populations Wayku, Kichwa Orellana and Sechura_Tallan. Posterior probabilities for the parameters were obtained by analysing $4x10^4$ simulations in an Approximate Bayesian Framework with the R package 'abc' (Csillery et al. 2012) and are shown in Table S4. Migration rates were set to 0 for 10^4 simulations, that were analysed separately to estimate split times in the absence of migration. Migration rates and split times were co-estimated for the other $3x10^3$ simulations.

Runs of homozygosity and consanguinity

Runs of homozygosity (ROH) blocks were identified with PLINK with default settings (Purcell et al. 2007). We divided ROH in each individual into two categories, long ROH (>1.6 Mb) and short to intermediate ROH (<1.6 Mb), based on the classes defined by Pemberton et al. (2012). While Pemberton et al. used a model based approach for ROH detection, an observational approach such as the one implemented in PLINK was shown to be very consistent in the recovery of ROH (Ceballos et al. 2018). We calculated the summed total length of ROH for each bin category for each individual. Long ROH were then further divided for a total of six bin categories and resulting ROH profiles were considered to describe possible demographic scenarios (e.g., recent bottleneck), similar to the study of Schroeder et al. (2018), which also considered Native American populations.

Phasing and IBD analysis

BEAGLE v 5.0. (Browning and Browning 2007) was used to phase the data. Before phasing, invariant sites were removed and the data was split into single chromosomes. Identity by descent (IBD) blocks were inferred with refined IBD (Browning and Browning 2013). Three runs of phasing and IBD detection were performed for each chromosome. The runs were then merged and gaps were removed with the utility provided, allowing a maximum gap length of 0.6 cM and at most 1 genotype in an IBD gap that is inconsistent with IBD. Only blocks with a minimum length of 2cM and LOD score >3 were retained for the analysis, to avoid spurious calls and errors in block merging (Browning and Browning 2013). The number of shared IBD blocks between pairs of populations was adjusted for sample size, by dividing by the product of the number of individuals in population 1 and population 2 in the pairwise comparison. Population pairwise sharing was considered only when more than one IBD block was retrieved, to further filter out spurious population connections. For the intra-continental comparisons,

we considered fragments larger than 5cM, a threshold used in previous work that has found that shorter fragments are ubiquitously shared across the entire continent (Harris et al. 2018).

Dating admixture events

Dating of admixture events was performed via two approaches. For dating with MALDER (Loh et al. 2013), populations with low sample sizes and similar levels of admixture (as estimated with the supervised ADMIXTURE analysis) were combined, and outlier individuals with exceptionally high level of admixture were excluded from populations in which admixture was otherwise low or absent (Sechura, Cofán, Kamentsa - see Table S2). MALDER assesses the exponential decay of admixture-induced LD in a target group, allowing for multiple admixture events (in this case for African, European and Native American sources). We ran MALDER with Yoruba, Spanish and three Native American parental populations, following the f_3 analysis scheme. Substituting data from Italian or French populations for the Spanish reference population did not change any of the results, and therefore results are only shown with the latter. Only admixture cases supported by p value<0.05 and Z score>3 were considered. For each population and for each of the Native American parental groups that passed this filtering, the pair with the highest Z score was kept.

As a second approach we used RFMix (Maples et al. 2013) to estimate local European, African or Native American ancestry along individual chromosomes, and then applied wavelet-transform analysis to the output, and used the WT coefficients to infer time since admixture by comparing the results to simulations, as described previously (Pugach et al. 2011; Pugach et al. 2016).

Time in generations ago was converted to calendar years assuming a generation time of 29 years (Fenner 2005).

Data visualization and source code

All data visualization was performed in R using packages developed by Wickham (2009), Becker et al. (2018), Kahle and Wickham (2013), and in-house scripts. The full detail of the command line setup and R scripts can be found at https://github.com/chiarabarbieri/SNPs_HumanOrigins_Recipes/

Author contributions: C.B. designed the research and analyzed the data; C.B, R.B., and A.M.T-A. performed laboratory analysis; C.B., R.B., L.A., J.R.S., O.A., C.Z., A.A-C., R.S-O. and R.F. collected samples; I.P. performed admixture analysis; F.M. performed simulation analysis; R.G. and M.S. provided laboratory resources and reagents; P.H., K.K.S, M.S, I.P. and L.F-S supervised the research; C.B. wrote the paper with inputs from all the coauthors.

Acknowledgments

We thank all the study participants and fieldwork assistants from Colombia, Ecuador, Mexico and Peru for making this study possible. The study was supported by a Wenner-Gren postdoctoral grant (Gr. 9395) to C.B., by the University Research Priority Program of Evolution in Action of the University of

Zurich to C.B. and K.K.S., and by MEXT Kakenhi 18H05080 to K.K.S. L.F.S. was supported by a U.S. National Science Foundation grant (NSF: A15-0187-001). L.A. was supported by a graduate grant from COLCIENCIAS. We thank David Reich and collaborators for providing a formatted dataset of published data used for population comparisons, Adrian Pearce for historical contextualization of European and African admixture in Peru, Vladimir Bajić, Hiba Babiker, Cosimo Posth and Thiseas Christos Lamnidis for computational analysis assistance.

Figure Legends:

Fig. 1. Map showing the approximate sampling locations of the newly reported population samples from South America, together with the ADMIXTURE results for K= 8. On top of the ADMIXTURE plot, newly reported population samples (in boldface) are shown together with other Native American samples from the literature, similarly typed with the Human Origins Affymetrix array. Yoruba and Spanish were also included in the ADMIXTURE runs to visualize African and European admixture.

Fig. 2. Principal component analysis of the newly reported samples together with representative populations from North and South America. A. First and second dimension. B. First and third dimension. C. First and fourth dimension. PCA was run with a subset of 2,545 SNPs previously defined as ascertained with Karitiana (see Materials and Methods). Color legend corresponds to geographic grouping. Three Cocama-speaking individuals from the "LoretoMix" population are marked with a red asterisk in the first PCA panel and discussed in the section on IBD analysis.

Fig. 3. Distribution of ROH classes. ROH analyses are run on a pruned dataset of 232,755 SNPs to avoid tracts affected by linkage disequilibrium. Classes of ROH are identified following Pemberton et al. (2012). A: Proportion of small and large ROH classes for each individual. B: ROH length classes profiles per groups, showing the variance of total length of ROH per each individual, binned for six length classes.

Fig. 4. Results of the IBD sharing analysis. A. Symmetrical matrix of pairwise IBD blocks sharing, showing the total length and the number of occurrences adjusted by population size. Populations are ordered by ecoregion and color-coded as in Fig. 2. B. Map visualizing the connections between populations that share blocks with each other: thin yellow lines indicate the lowest levels of exchange, thick red lines the highest (adjusted for population size). Only blocks larger than 5 cM are considered.

Fig. 5. Admixture dates between European and African sources. Estimates of admixture are calculated with the MALDER and WAVELETS methods. Dates are expressed in generations ago and converted to calendar years using a generation time of 29 years.

References

Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Hooshiar Kashani B, Battaglia V, Grugni V, Angerhofer N, Rogers MP, et al. 2013. Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proc. Natl. Acad. Sci. U. S. A. 110:14308–14313.

Acuña-Alonzo V, Flores-Dorantes T, Kruit JK, Villarreal-Molina T, Arellano-Campos O, Hünemeier T, Moreno-Estrada A, Ortiz-López MG, Villamil-Ramírez H, León-Mimila P, et al. 2010. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Hum. Mol. Genet. 19:2877–2885.

Adelaar WFH, Muysken PC. 2004. The Languages of the Andes. Cambridge, UK: Cambridge University Press.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Arias L, Barbieri C, Barreto G, Stoneking M, Pakendorf B. 2018. High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia. Am. J. Phys. Anthropol. 165:238–255.

Arias L, Schroeder R, Huebner A, Barreto G, Stoneking M, Pakendorf B. 2018. Cultural Innovations influence patterns of genetic diversity in Northwestern Amazonia. Mol. Biol. Evol. 35:2719–2735.

Arrelucea Barrantes M, Cosamalón Aguilar JA. 2015. La Presencia Afrodescendiente en el Perú. Siglos XVI-XX. Lima, Peru: Ministerio de Cultura.

Barbieri C, Heggarty P, Yang Yao D, Ferri G, De Fanti S, Sarno S, Ciani G, Boattini A, Luiselli D, Pettener D. 2014. Between Andes and Amazon: The genetic profile of the Arawak-speaking Yanesha. Am. J. Phys. Anthropol. 155:600–609.

Barbieri C, Sandoval JR, Valqui J, Shimelman A, Ziemendorff S, Schröder R, Geppert M, Roewer L, Gray R, Stoneking M, et al. 2017. Enclaves of genetic diversity resisted Inca impacts on population history. Sci. Rep. 7:17411.

Becker R, Wilks A, Brownrigg R, Minka T, Deckmyn. A. 2018. maps: Draw Geographical Maps. R package version 3.3.0. https://CRAN.R-project.org/package=maps.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics 32:2817–2823.

Bisso-Machado R, Bortolini MC, Salzano FM. 2012. Uniparental genetic markers in South Amerindians. Genet. Mol. Biol. 35:365–387.

Bolnick DA, Raff JA, Springs LC, Reynolds AW, Miró-Herrans AT. 2016. Native American genomics and population histories. Annu. Rev. Anthropol. 45:319–340.

Brandini S, Bergamaschi P, Fernando Cerna M, Gandini F, Bastaroli F, Bertolini E, Cereda C, Ferretti L, Gómez-Carballa A, Battaglia V, et al. 2018. The Paleo-Indian entry into South America according to mitogenomes. Mol. Biol. Evol. 35:299–311.

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459–471.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097.

Cabana GS, Lewis CM, Tito RY, Covey RA, Cáceres AM, De La Cruz AF, Durand D, Housman G, Hulsey BI, Iannacone GC, et al. 2015. Population genetic structure of traditional populations in the Peruvian Central Andes and implications for South American population history. Hum. Biol.

86:147-166.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. 19:220–234.

Cerrón-Palomino R. 2003. Lingüística Quechua. 2nd ed. Cuzco, Peru: Bartolomé de Las Casas.

Chacón-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, Quinto-Sánchez M, Gómez-Valdés J, Everardo Martínez P, Villamil-Ramírez H, et al. 2018. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. Nat. Commun. 9:5388.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4:7.

Church W, Von Hagen A. 2008. Chachapoyas: Cultural development at an Andean cloud forest crossroads. In: Silverman H, Isbell WH, editors. The Handbook of South American Archaeology. New York: Springer. p. 903–926.

Clement CR, de Cristo-Araújo M, D'Eeckenbrugge GC, Pereira AA, Picanço-Rodrigues D. 2010. Origin and domestication of native Amazonian crops. Diversity 2:72–106.

Csillery K, Francois O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3:475–479.

D'Altroy TN. 2014. The Incas. New York: John Wiley & Sons.

Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. Science 300:597–603.

Dillehay TD, Goodbred S, Pino M, Vásquez Sánchez VF, Tham TR, Adovasio J, Collins MB, Netherly PJ, Hastorf CA, Chiou KL, et al. 2017. Simple technologies and diverse food strategies of the Late Pleistocene and Early Holocene at Huaca Prieta, Coastal Peru. Sci. Adv. 3:e1602778.

Dixon EJ. 2013. Late Pleistocene colonization of North America from Northeast Asia: New insights from large-scale paleogeographic reconstructions. Quat. Int. 285:57–67

Dixon RMW, Aikhenvald AY. 1999. The Amazonian languages. Cambridge: Cambridge University Press.

Epps P. 2009. Language Classification, Language Contact, and Amazonian Prehistory. Lang. Linguist. Compass 3:581–606.

Fehren-Schmitz L, Haak W, Mächtle B, Masch F, Llamas B, Cagigao ET, Sossna V, Schittek K, Isla Cuadrado J, Eitel B, et al. 2014. Climate change underlies global demographic, genetic, and cultural transitions in pre-Columbian southern Peru. Proc. Natl. Acad. Sci. U. S. A. 111:9443–9448.

Fehren-Schmitz L, Reindel M, Cagigao ET, Hummel S, Herrmann B. 2010. Pre-Columbian population dynamics in coastal southern Peru: A diachronic investigation of mtDNA patterns in the Palpa region by ancient DNA analysis. Am. J. Phys. Anthropol. 141:208–221.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. 128:415–423.

de Filippo C, Bostoen K, Stoneking M, Pakendorf B. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc. R. Soc. B-Biological Sci. 279:3256–3263.

Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Mol. Biol. Evol. 20:1682–1691.

Goldberg A, Mychajliw AM, Hadly EA. 2016. Post-invasion demography of prehistoric humans in South America. Nature 532:232–235.

Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, Kenny EE, Gignoux CR, Maples BK, Guiblet W, et al. 2013. Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. 9: e1004023

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211.

Harris DN, Song W, Shetty AC. Lavano KS, Caceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanches C, et al. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc Natl Acid Sci USA 1115:E6526-E6536.

Heckenberger M, Neves EG. 2009. Amazonian Archaeology. Annu. Rev. Anthropol. 38:251–266.

Heggarty P. 2008. Linguistics for Archaeologists: a Case-study in the Andes. Cambridge Archaeol. J. 18:35–56.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet. 11:e1005602.

Hornborg A, Gassn R, Heckenberger M, Hill J, Neves E, SantosGranero F. 2005. Ethnogenesis, regional integration, and ecology in prehistoric Amazonia: toward a system perspective. Curr. Anthropol. 46:589–620.

Isbell WH. 2008. Wari and Tiwanaku: International Identities in the Central Andean Middle Horizon. In: Silverman H, Isbell W, editors. The Handbook of South American Archaeology. New York: Springer. p. 731–759.

Kahle D, Wickham H. 2013. ggmap : Spatial Visualization with ggplot2. R J. 5:144–161.

Kirin M, McQuillan R, Franklin CS, Campbell H, Mckeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. PLoS One 5:e13996.

de la Fuente C, Ávila-Arcos MC, Galimany J, Carpenter ML, Homburger JR, Blanco A, Contreras P, Cruz Dávalos D, Reyes O, San Roman M, et al. 2018. Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. Proc. Natl. Acad. Sci. U.S.A. 115:E4006–E4012.

Lathrap DW. 1970. The upper Amazon. London: Thames & Hudson.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413.

Lipson M, Loh P-R, Patterson N, Moorjani P, Ko Y-C, Stoneking M, Berger B, Reich D. 2014. Reconstructing Austronesian population history in Island Southeast Asia. Nat. Commun. 5:4689.

Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohlach A, et al. 2016. Ancient mitochondrial DNA provides high-resolution timescale of the peopling of the Americas. Sci. Adv. 2:1–10.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics 193:1233–1254.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93:278–288.

Mezzavilla M, Geppert M, Tyler-Smith C, Roewer L, Xue Y. 2015. Insights into the origin of rare haplogroup C3* Y chromosomes in South America from high-density autosomal SNP genotyping. Forensic Sci. Int. Genet. 15:115–120.

Michael L. 2014. On the Pre-Columbian origin of proto-omagua-kokama. J. Lang. Contact 7:309–344.

Mooney JA, Huber CD, Service S, Sul JH, Marsden CD, Zhang Z, Sabatti C, Ruiz-Linares A, Bedoya G, Costa Rica/Colombia Consortium for Genetic Investigation of Bipolar Endophenotypes, et al. 2018. Understanding the Hidden Complexity of Latin American Population Isolates. Am. J. Hum. Genet. 103:707–726.

Moreno-Mayar VJ, Vinner L, Damgaard P de B, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, et al. 2018. Early human dispersals within the Americas. Science. 362:eaav2621.

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. Am. J. Hum. Genet. 91:809–822.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. 91:275–292.

Pemberton TJ, DeGiorgio M, Rosenberg NA. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. G3 (Bethesda). 3:891–907.

Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E, et al. 2018. Reconstructing the deep population history of Central and South America. Cell 175:1185-1197.

Poznik GD. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv:088716.

Pugach I, Duggan AT, Merriwether DA, Friedlaender FR, Friedlaender JS, Stoneking M. 2018. The gateway from near into remote oceania: New insights from genome-wide data. Mol. Biol. Evol. 35:871–886.

Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, Pakendorf B. 2016. The complex admixture history and recent southern origins of Siberian populations. Mol. Biol. Evol. 33:1777–1795.

Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol. 12:R19.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Quilter J. 2013. The Ancient Central Andes. London and New York: Routledge.

Rademaker K, Hodgins G, Moore K, Zarrillo S, Miller C, Bromley G, Leach P, Reid D, Yépez Álvarez W, Sandweiss D. 2014. Paleoindian settlement of the high-altitude Peruvian Andes. Science. 24:466–469.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. 2014. Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. Nature 505:87–91.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A, et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 349:aab3884.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al. 2012. Reconstructing Native American population history. Nature 488:370–374.

Renfrew C, Bellwood P. 2002. Examining the farming/language dispersal hypothesis. Cambridge: McDonald Institute for Archaeological Research.

Rothhammer F, Dillehay TD. 2009. The late Pleistocene colonization of South America: an interdisciplinary perspective. Ann. Hum. Genet. 73:540–549.

Sandoval JR, Lacerda DR, Acosta O, Jota MS, Robles-Ruiz P, Salazar-Granara A, Vieira PPR, Paz-y-Miño C, Fujita R, Santos FR, et al. 2016. The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations. Ann. Hum. Genet. 80:88–101.

Sandoval JR, Lacerda DR, Jota MS, Elward R, Acosta O, Pinedo D, Danos P, Cuellar C, Revollo S, Santos FR, et al. 2018. Genetic ancestry of families of putative Inka descent. Mol. Genet. Genomics: 293:873–881

Sandoval JR, Salazar-Granara A, Acosta O, Castillo-Herrera W, Fujita R, Pena SD, Santos FR. 2013. Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. J. Hum. Genet. 58:627–634.

Sandweiss DH. 2008. Early fishing societies in western South America. In: Silverman H, Isbell W, editors. Handbook of South American Archaeology. New York: Springer. p. 145–156.

Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Mörseburg A, Johnson JR, Potter A, et al. 2018. Ancient human parallel lineages within North America contributed to a coastal expansion. Science. 360:1024–1027.

Schroeder H, Sikora M, Gopalakrishnan S, Cassidy LM, Maisano Delser P, Sandoval Velasco M, Schraiber JG, Rasmussen S, Homburger JR, Ávila-Arcos MC, et al. 2018. Origins and genetic legacies of the Caribbean Taino. Proc. Natl. Acad. Sci. U.S.A 115:2341-2346.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. 2015. Genetic evidence for two founding populations of the Americas. Nature. 525:104–108.

Socolow SM. 2015. Women, Marriage, and Family. In: The Women of Colonial Latin America. Cambridge: Cambridge University Press.

de Souza JG, Schaan DP, Robinson M, Barbosa AD, Aragão LEOC, Marimon BH, Marimon BS, da Silva IB, Khan SS, Nakahara FR, et al. 2018. Pre-Columbian earth-builders settled along the entire southern rim of the Amazon. Nat. Commun. 9:1125.

Staab PR, Zhu S, Metzler D, Lunter G. 2015. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. Bioinformatics 31:1680–1682.

Stanish C. 2001. The origin of state societies in South America. Annu. Rev. Anthropol. 30:41–64.

Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, et al. 2007. Beringian standstill and spread of Native American founders. PLoS One 2:e829.

Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, Rickards O, Tyler-Smith C, Pena SDJ, Santos FR. 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am. J. Hum. Genet. 68:1485–1496.

Valverde G, Romero MIB, Espinoza IF, Cooper A, Fehren-Schmitz L, Llamas B, Haak W. 2016. Ancient DNA analysis suggests negligible impact of the Wari empire expansion in Peru's central coast during the Middle Horizon. PLoS One 11:e0155508.

Venables WN, Ripley BD. 2002. MASS: modern applied statistics with S. New York: Springer.

Wang S, Lewis Jr. CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, et al. 2007. Genetic Variation and Population Structure in Native Americans. PLoS Genet 3:e185.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 38:1358-1370.

Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, Kronenberg F, Salas A, Schönherr S. 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 44:W58–W63.

Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.